High Performance Global File Systems
Easy Data Management in Supercomputer Grids

Distributed European Infrastructure for Supercomputing Applications

Andreas Schott (schott@rzg.mpg.de)

Overview

Distributed European Infrastructure for Supercomputing Applications

- Motivation / Choices
- GPFS / MC-GPFS
- DEISA's Implementation and Status

Motivation for Global File Systems

Distributed European Infrastructure for Supercomputing Applications

Advantages

- Simple access
- Standard commands
- No special data preparation
- No re-writing of jobs and binaries
- Everything everywhere at any time

Issues

- Network stability
- Latency
- Performance
- Availability

Available Choices

Distributed European Infrastructure for Supercomputing Applications

- (Open)AFS

- GFS
- PVFS
- OCFS

- NFS
- NFS4
- Lustre

- MC-GPFS

General Concepts of MC-GPFS

Distributed European Infrastructure for Supercomputing Applications

MC-GPFS = Multiple Cluster General Parallel File System

- available for all HPC architectures in DEISA
- servers available for AIX and Linux

Principle Structure

- distributed – shared – striped
- kernel add-on for file system
- block oriented data transfer

Features achieved

- shared and high performance access
- safe and secure data
- high administrative flexibility

General Concepts of MC-GPFS

Distributed European Infrastructure for Supercomputing Applications

Technical Aspects

- each site with its own servers possible
  - local disk space locally administered
- scalability and high performance access by inherent parallelism
  - easy extensible
- file consistency by sophisticated token management
  - high recoverability and increased data availability
- simplified storage management
  - storage pools, file sets
- simplified administration
  - globally acting commands

General Concepts of MC-GPFS

Distributed European Infrastructure for Supercomputing Applications

Security Aspects

- separate network communication for administration possible
- remote security
  - authenticated remote access for servers
    - mount and/or data with SSL-keys
- easy root-mapping
- easy no-suid functionality
- userid mapping for remote access via interfaces

General Concepts of MC-GPFS

Distributed European Infrastructure for Supercomputing Applications

Access and Availability

- transparent access
  - no special data transfer commands required
- global visibility inside DEISA
- extended access rights
- no single point of failure communication
- delegated locking and other communication

Summary of MC-GPFS

Distributed European Infrastructure for Supercomputing Applications

Local and Remote High Performance Access

- high parallelism in data and file access
- very large file and file system support

High Availability

- each site with its own servers
- redundant access path
- simply extensible and scalable
- striped data
- parallel access path

Advantages of GPFS (admin)

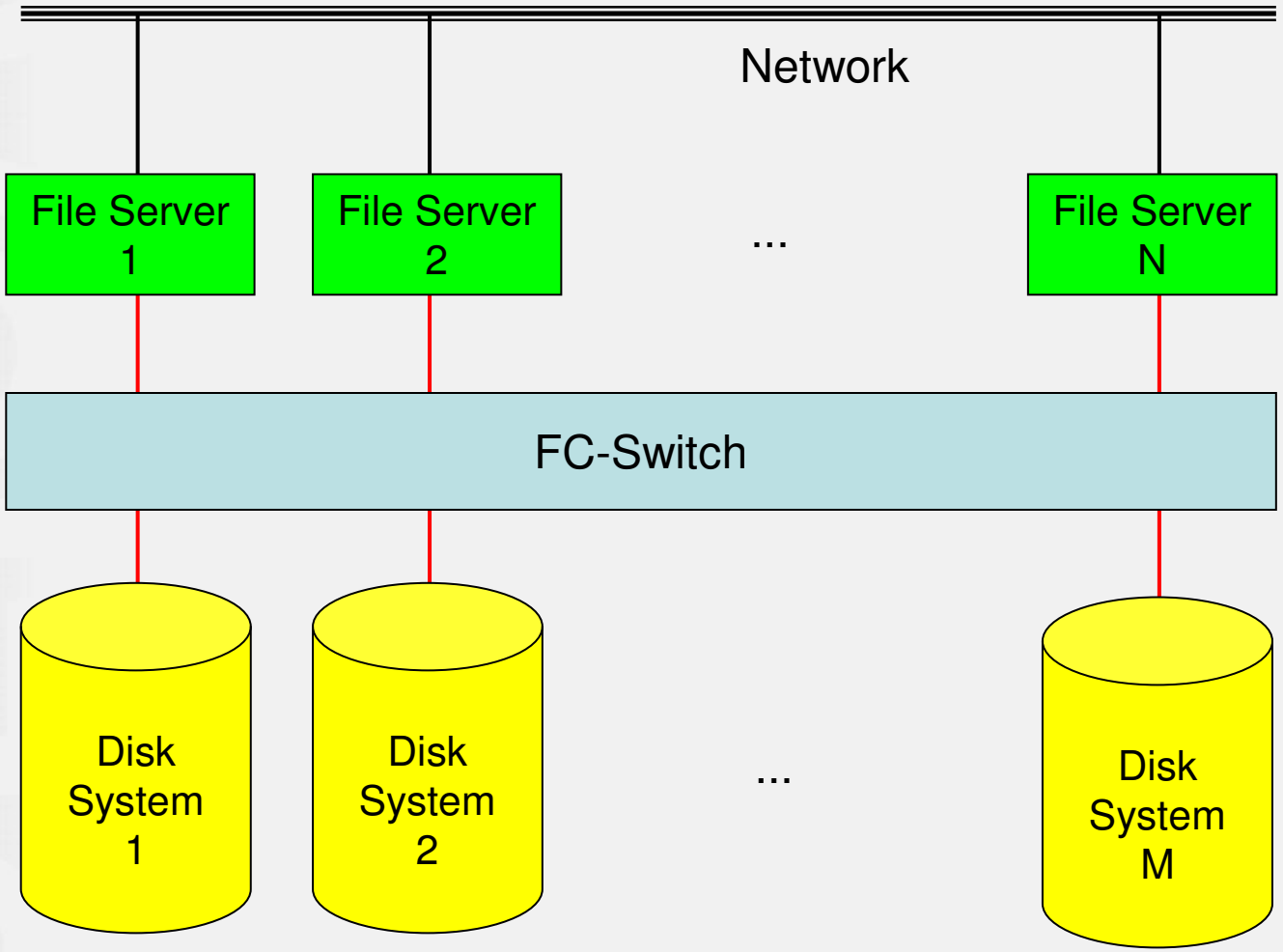
- Easy Management
- Easy Extensibility
- High Performance
- Security Features
- Add-On Features like HSM Functionality

Advantages of GPFS (user)

Distributed European Infrastructure for Supercomputing Applications

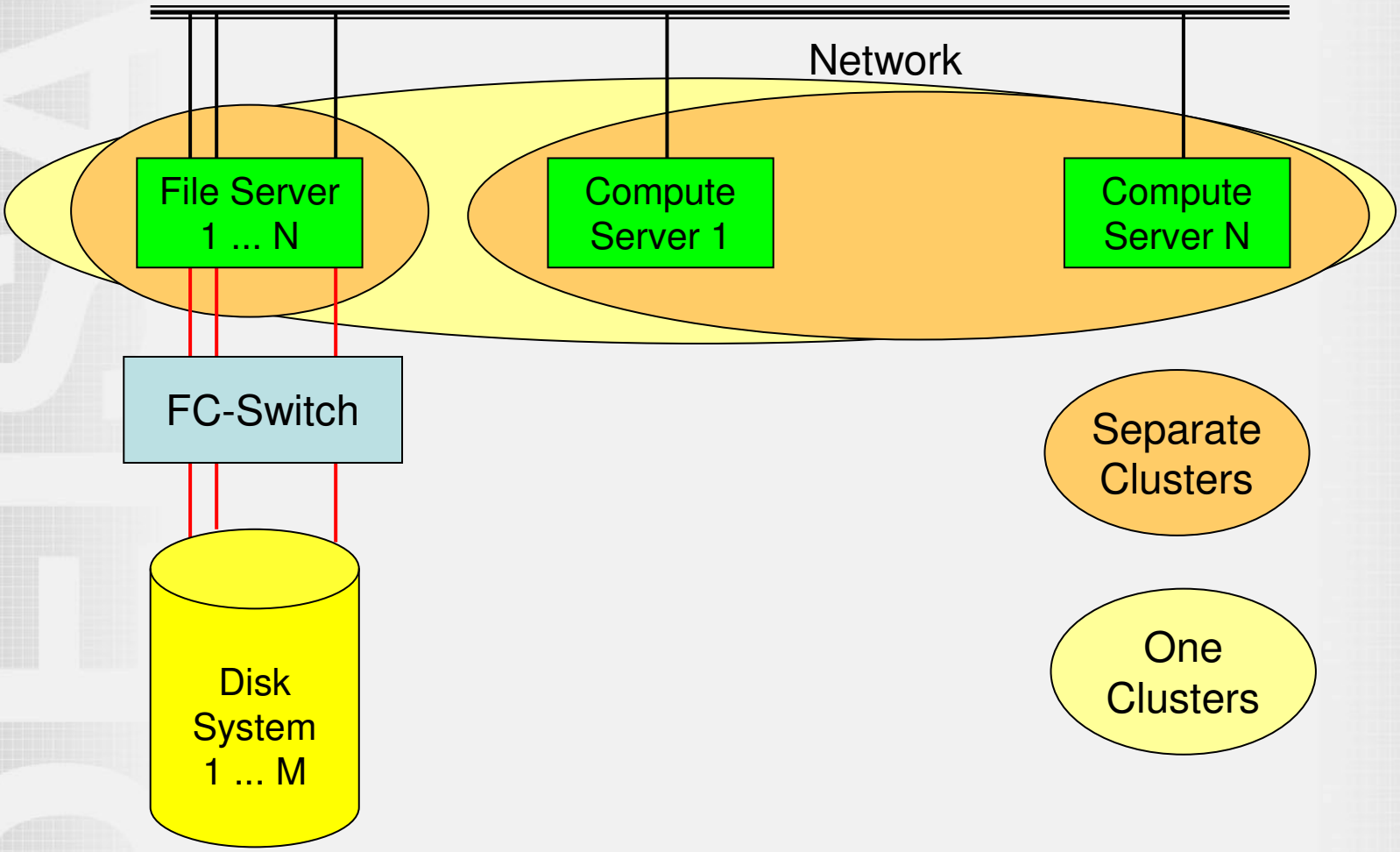
- Standard Access Methods
  - Transparent Access
- Data globally visible
  - No special actions for data transfer required
- Simplicity
- Extended Access Right Features
- Add-On Features like HSM Functionality

Local GPFS File Servers



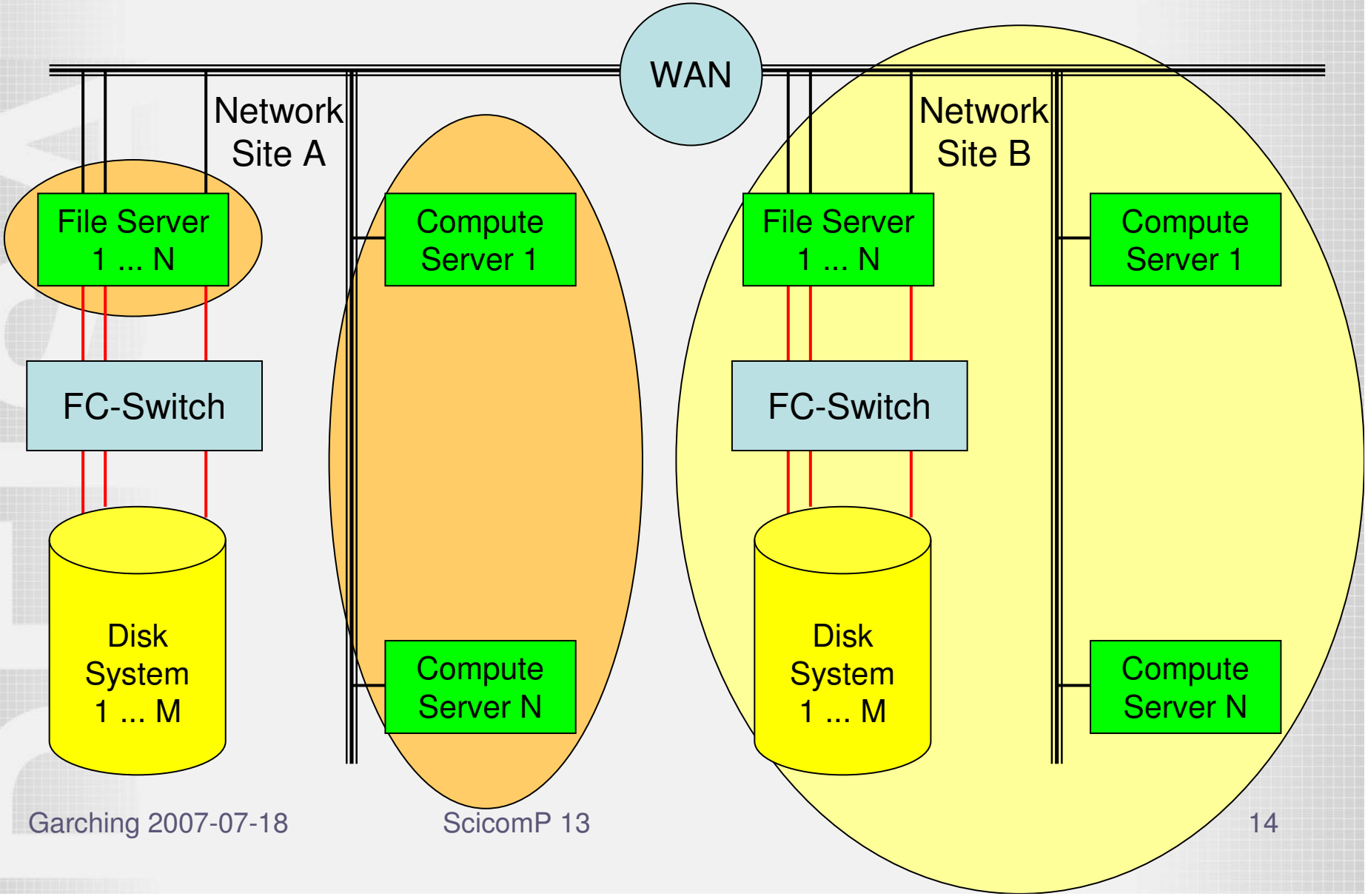
Local GPFS Access

Distributed European Infrastructure for Supercomputing Applications



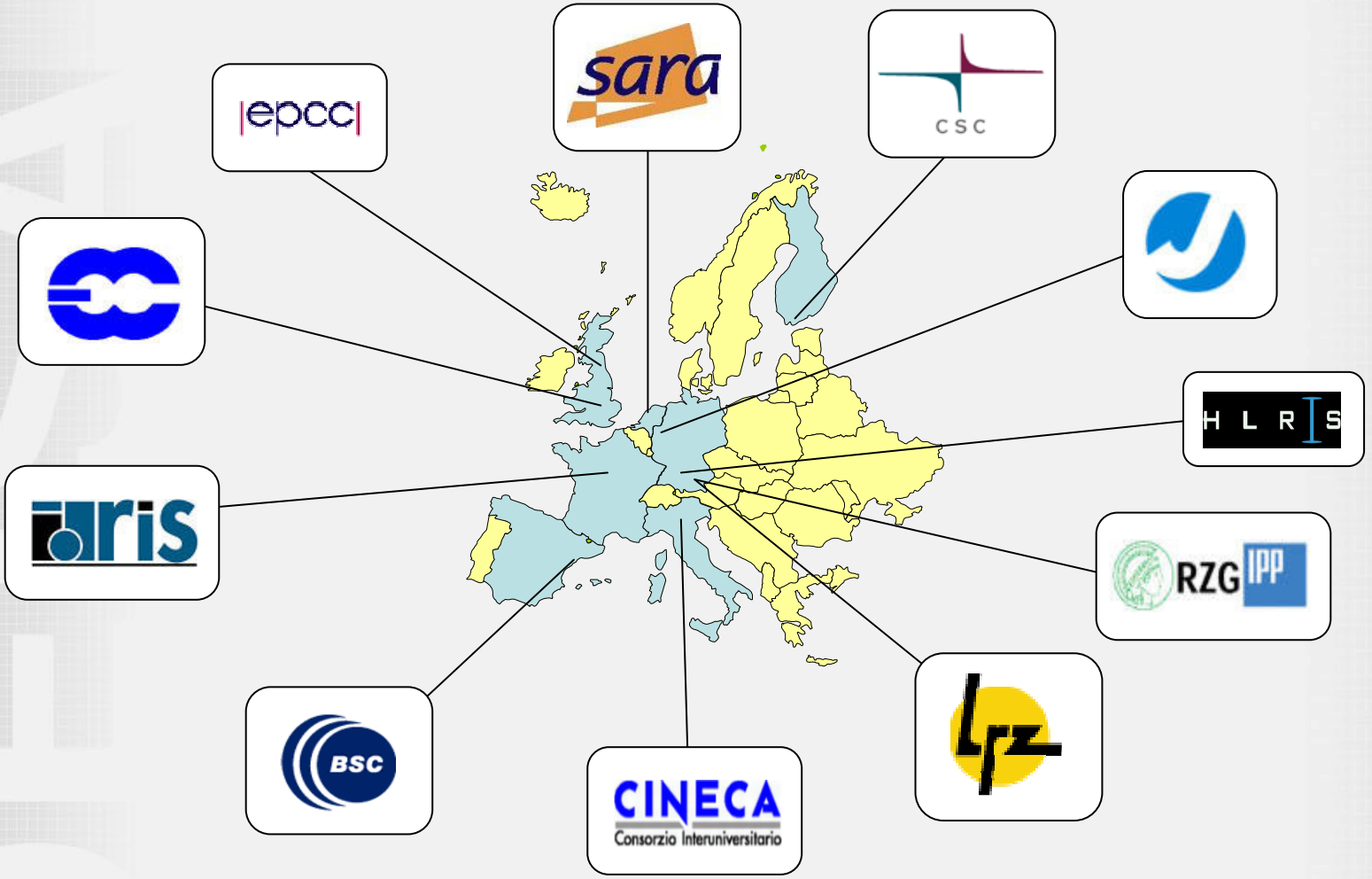
Remote GPFS Access

Distributed European Infrastructure for Supercomputing Applications



DEISA Partners

Distributed European Infrastructure for Supercomputing Applications



Aims of DEISA

Distributed European Infrastructure for Supercomputing Applications

Providing HPC resources to the Scientific Community

- Offering an add-on value to local facilities
  - optimal hardware selection
  - easy usability
  - transparent data access

Achievement of these Aims

- common network structure
- using internal features of job schedulers
- additional middleware for easy access (e.g. UNICORE)
- global file system in a network of trust

MC-LoadLeveler in DEISA

Distributed European Infrastructure for Supercomputing Applications

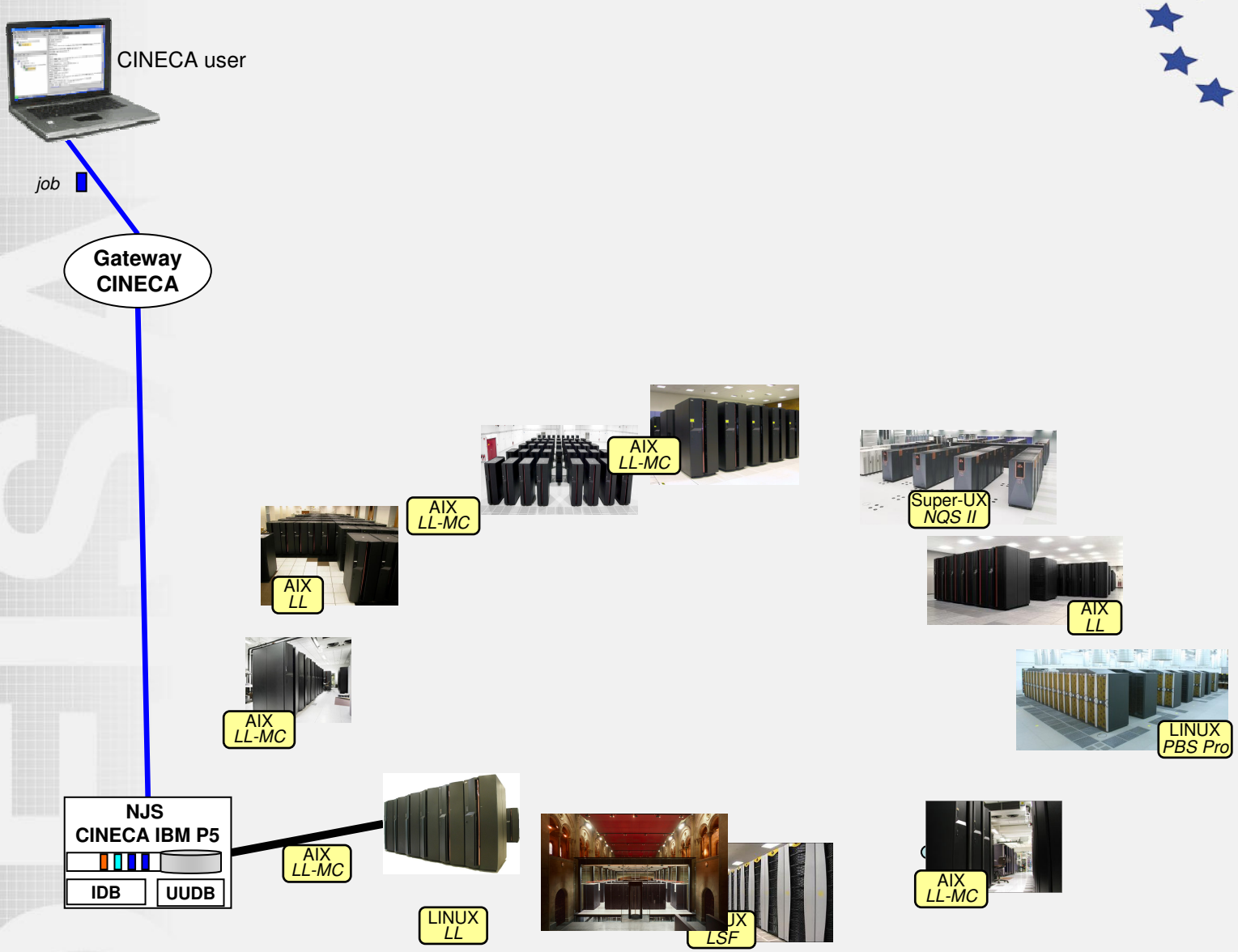
Implementation

- Environment Variables for DATA
- Modules
- Local Home Directories
- Job Movement (Filters)

Caveats

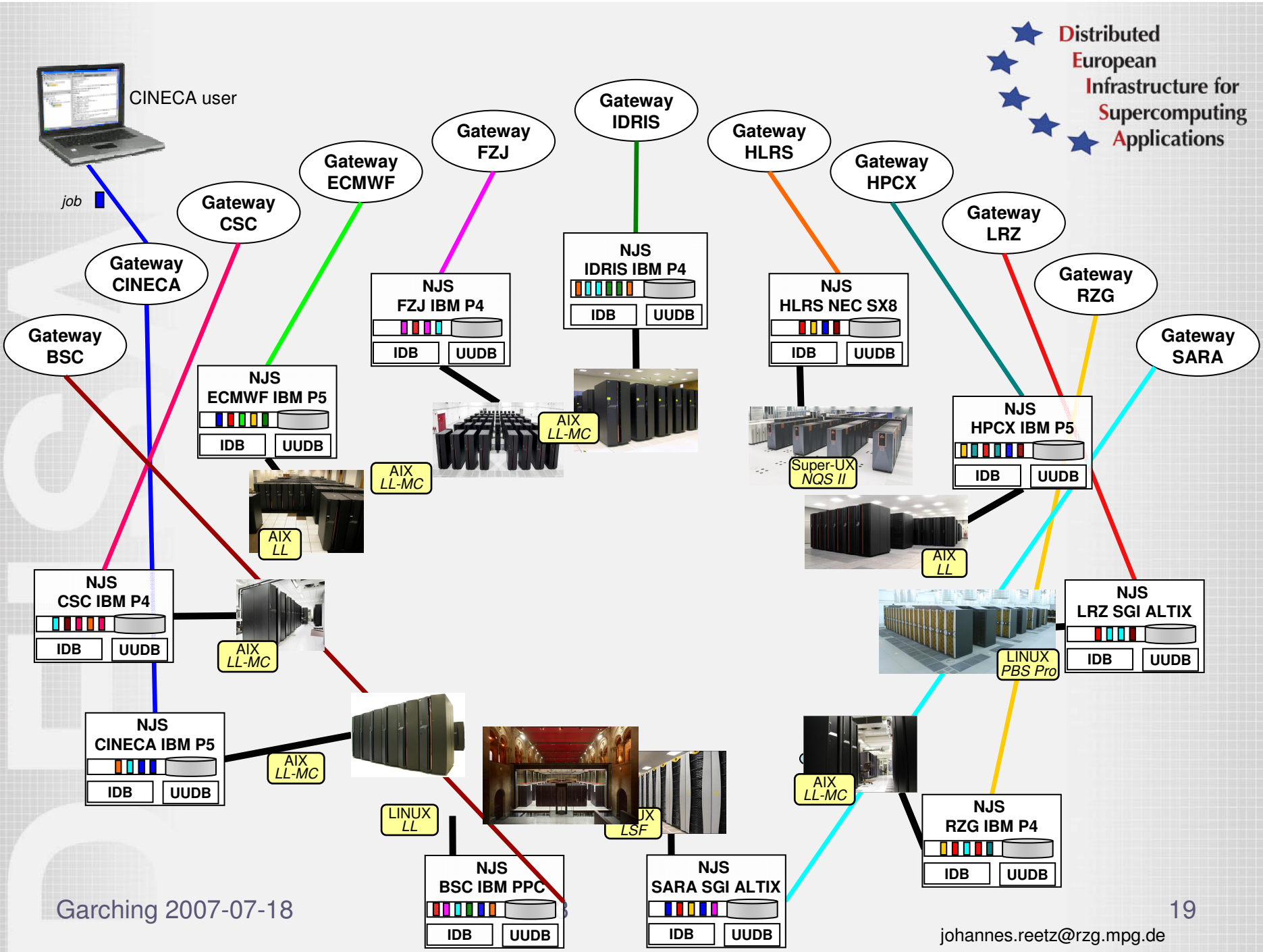
- Path Unification
- Treatment of HSM
- Data Availability

Pre- and Post-processing

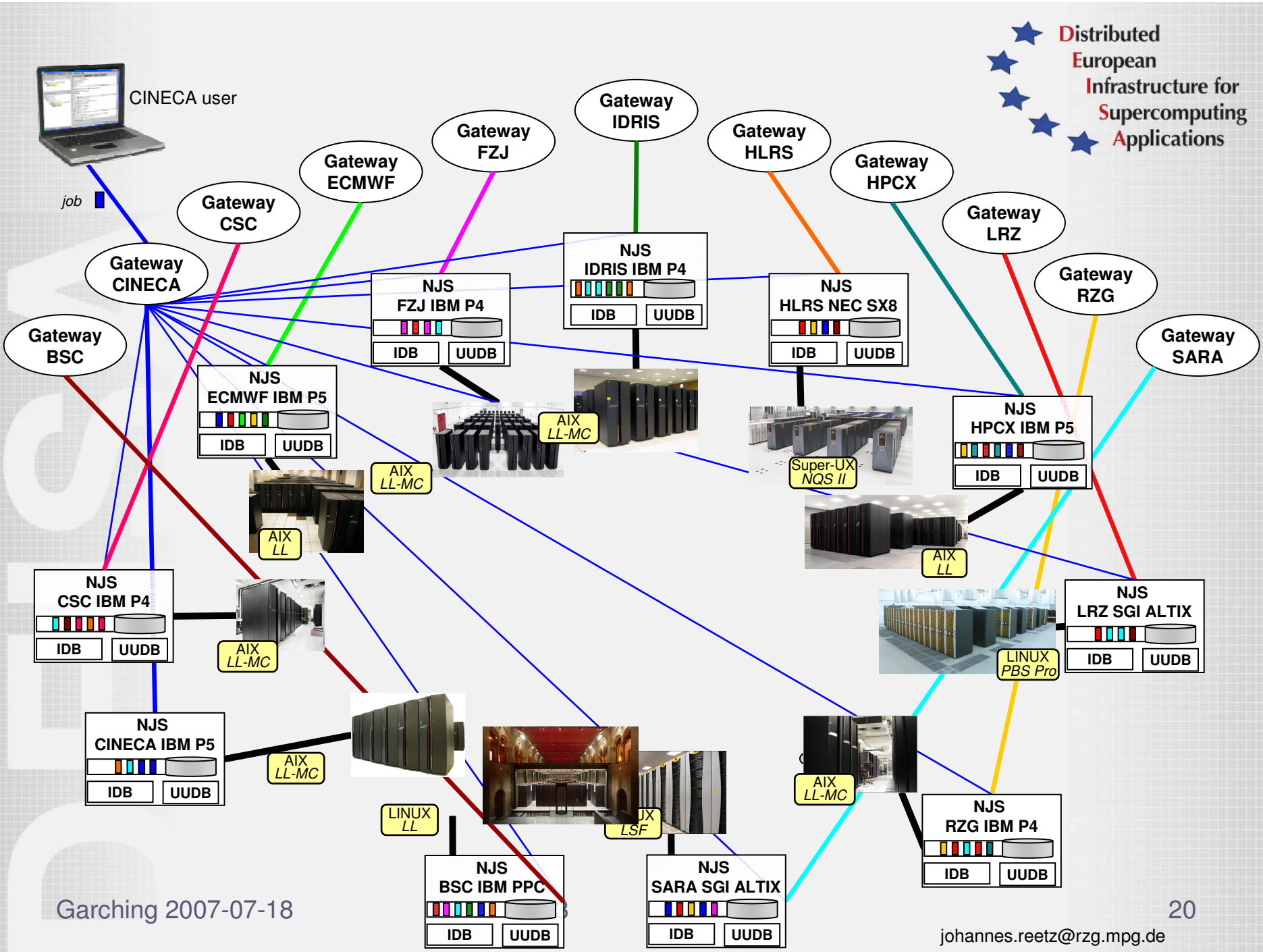


Distributed European Infrastructure for Supercomputing Applications

johannes.reetz@rzg.mpg.de



johannes.reetz@rzg.mpg.de

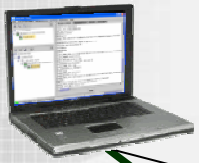


johannes.reetz@rzg.mpg.de

Globus Installation at RZG

Distributed European Infrastructure for Supercomputing Applications

D-GRID user



GLOBUS client tools
grid-proxy-init
globusrun-ws
globus-url-copy
gsissh

internet

gsissh

grid gateway (job submission host)

gridftp frontend *2811* *(user mode)*
gridftp backend *(root)*

globus container 8443
DMZ *firewall inbound ports (8443,20000-25000)*

(fork), LRMS client
GPFS available
grid-mapfile: (DN → D-GRID username)

head node (e.g., for code development and testing)
gsisshd 2222

LRMS client
full DEISA CPE available

LRMS (node hosting the LoadLeveler master)
Local Resource Management System (IBM LoadLeverer)

grid gateway
gg.rzg.mpg.de

Linux

DMZ

RFT

GPFS

Cluster compute nodes (IBM P5)

AIX

p5io3.rzg.mpg.de
(head node)

LRMS (master)

IO node

high performance switch

intranet

postgres DB

disk system

Johannes.Reetz@rzg.mpg.de

GPFS Configuration in DEISA

Distributed European Infrastructure for Supercomputing Applications

Each AIX-site provides its own server

Some non-AIX-sites will provide servers based on Linux

RZG hosts disk space for non-AIX-sites without servers

RZG provides HSM-functionality on GPFS

locally disk space performs like local disk space

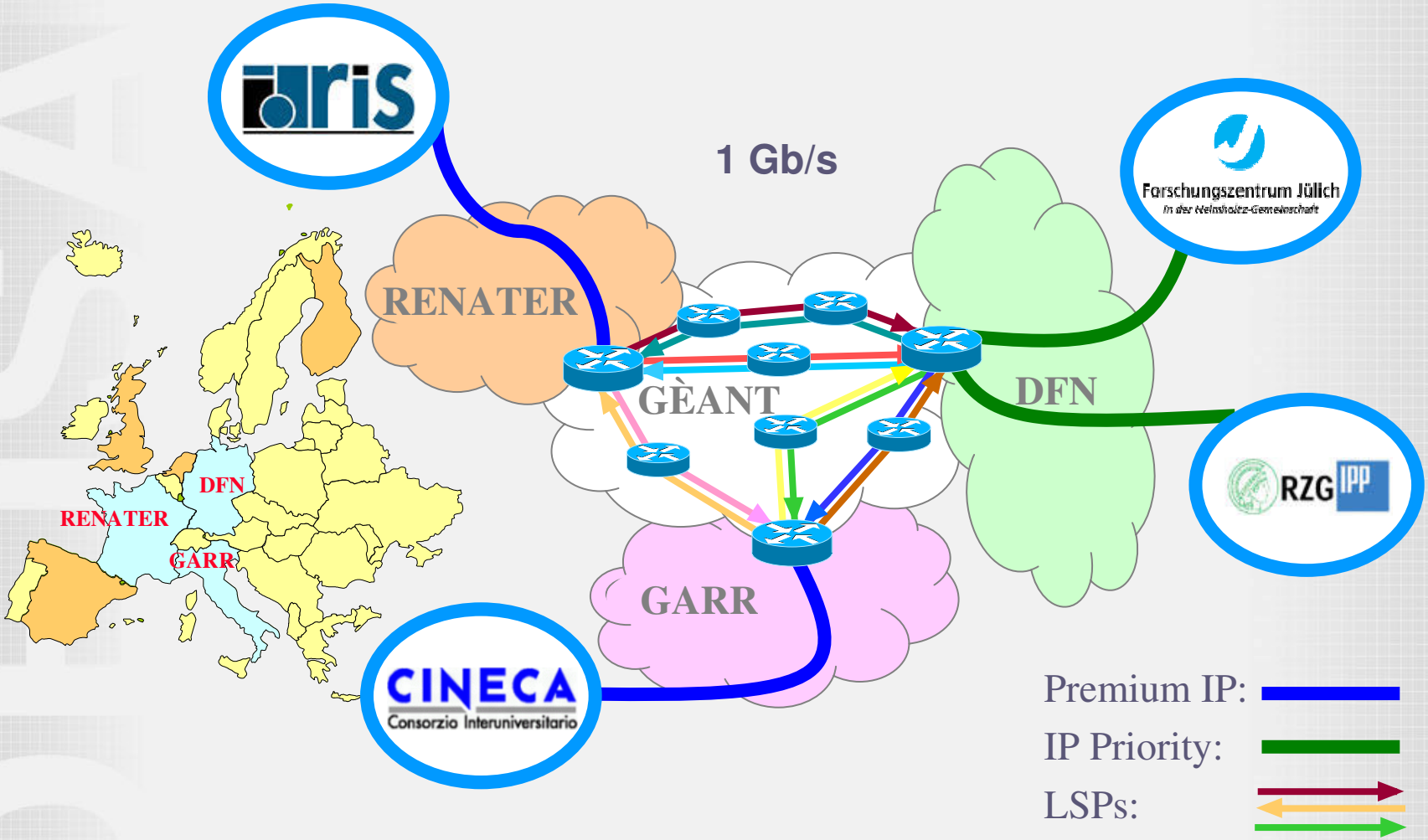
total of more than 30 TB

wide area network connection with 10GBit/s (mostly)

remotely disk space no longer limited by network

DEISA „proof of concept“ phase

Distributed European Infrastructure for Supercomputing Applications



Evolution of GPFS in DEISA

Distributed European Infrastructure for Supercomputing Applications

FZJ (DE)
Power4
AIX

October 2004

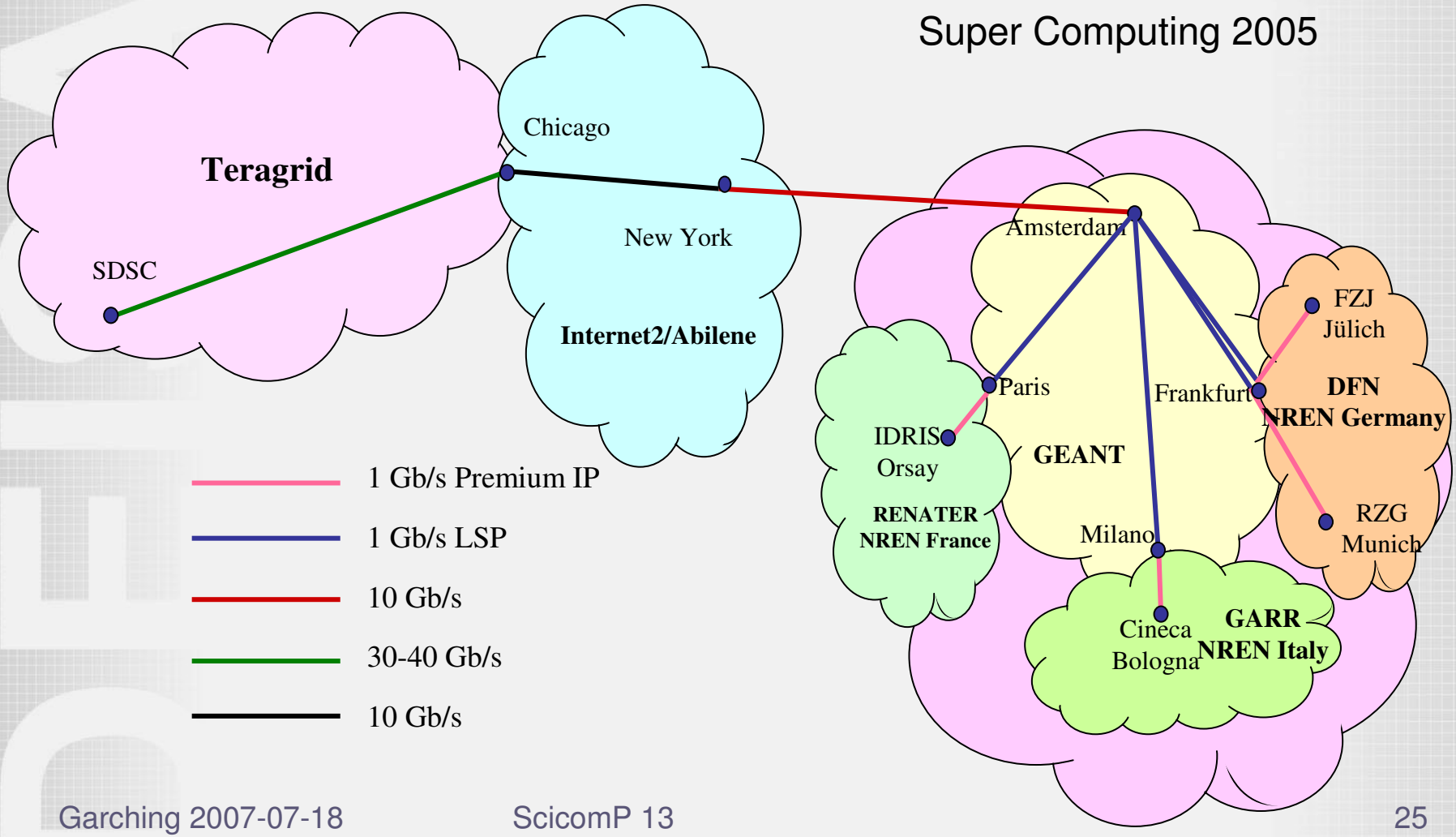
IDRIS (FR)
Power4
AIX

RZG (DE)
Power4
AIX

CINECA (IT)
Power5
AIX

DEISA – TeraGrid Connection

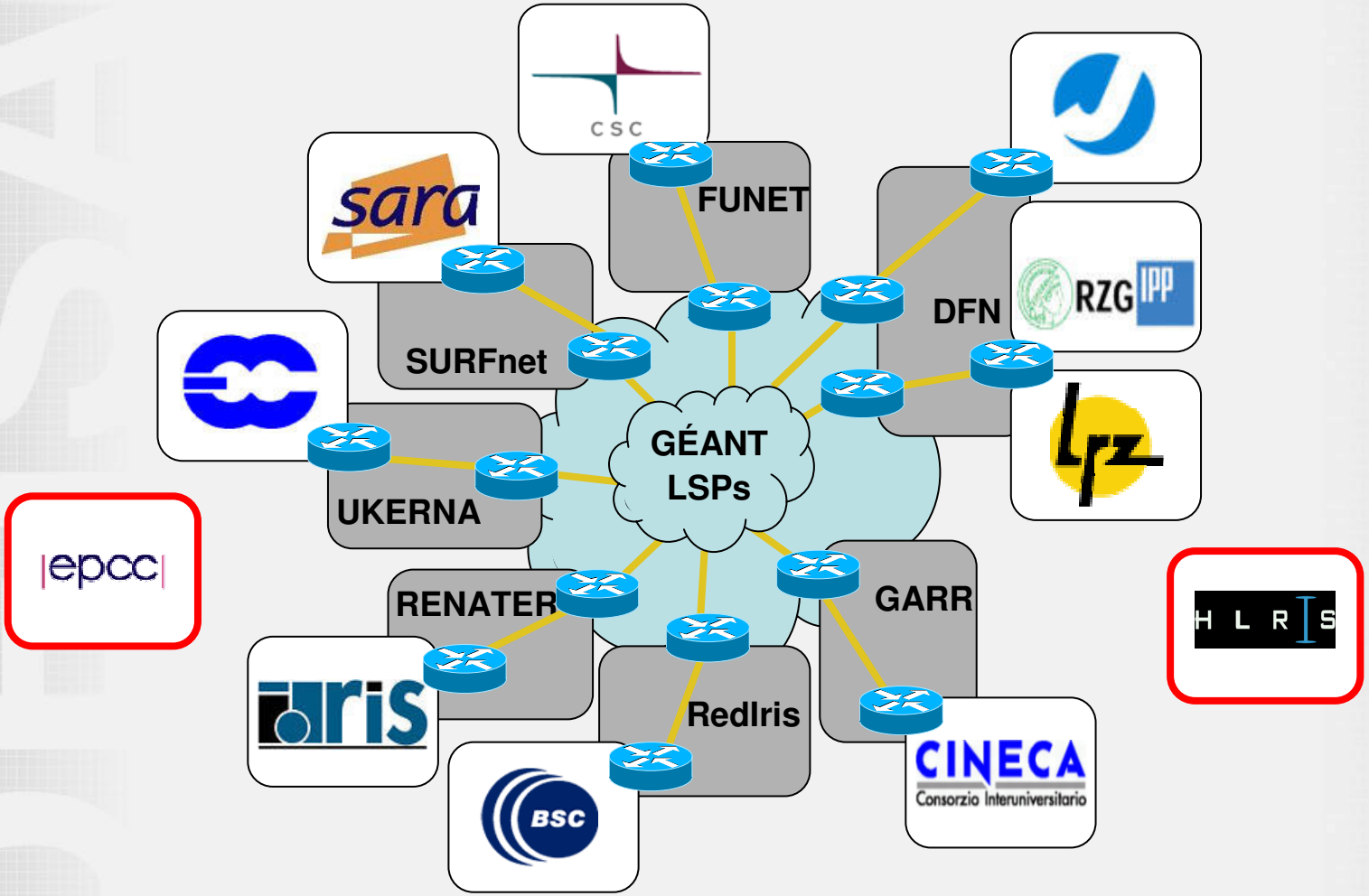
Distributed European Infrastructure for Supercomputing Applications



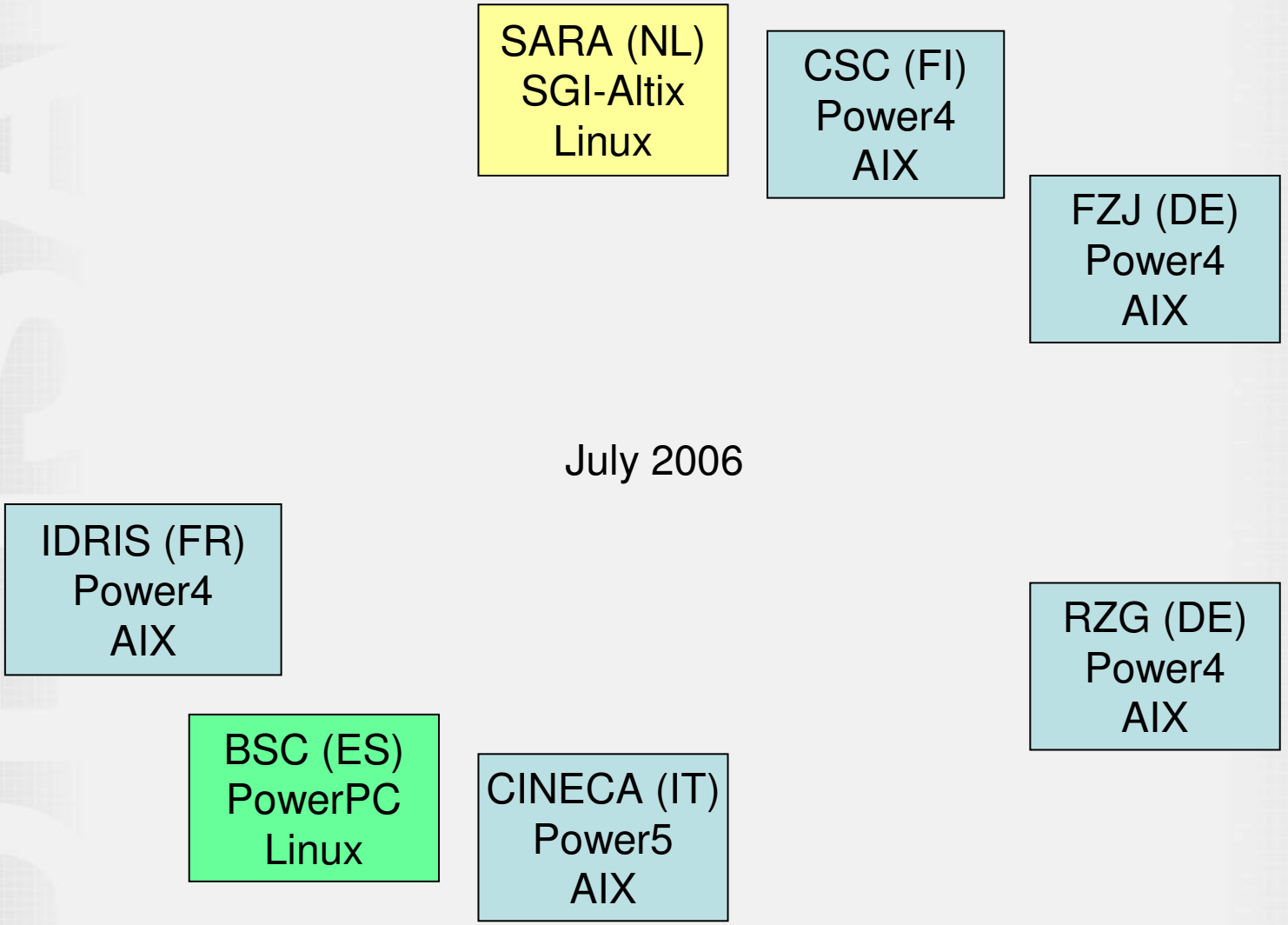
R.Niederberger@fz-juelich.de

DEISA 1 Gb/s network infrastructure

Distributed European Infrastructure for Supercomputing Applications



Evolution of GPFS in DEISA



Upgrade of Multiple Cluster GPFS

Problems with GPFS 2.3

- Initial MC-functionality not inherently integrated
- Each-to-Any communication required
- Limitation of participating nodes

Advantages of GPFS 3.1

- Better Multi-Cluster Support
- Better Encapsulation by possible use of private addresses
- Higher Independence between sites
- Higher Stability
- Better Performance

Evolution of GPFS in DEISA

Distributed European Infrastructure for Supercomputing Applications

CSC (FI)
Power4
AIX

FZJ (DE)
Power4
AIX

ECMWF (GB)
Power5+
AIX

February 2007

IDRIS (FR)
Power4
AIX

RZG (DE)
Power4
AIX

LRZ (DE)
SGI-Altix
Linux

CINECA (IT)
Power5
AIX

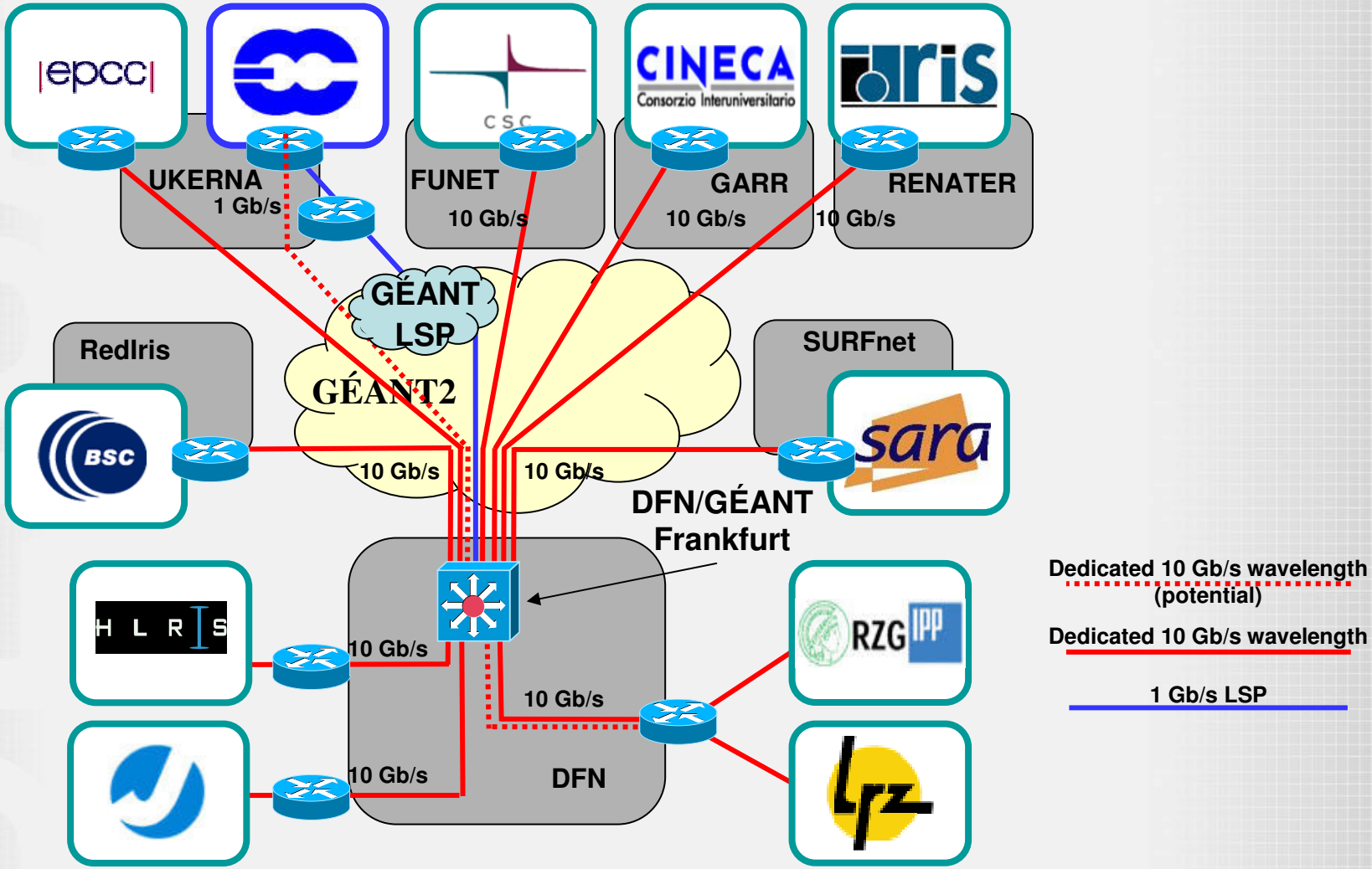
Status of Multiple Cluster GPFS

Distributed European Infrastructure for Supercomputing Applications

Site	File-server	Storage	Compute-CPUs	TFlops	Memory
CINECA	2	2 TB	480 Power5 (1.9 GHz)	2.6	1152 GB
CSC	2	2 TB	512 Power4 (1.1 GHz)	2.2	672 GB
ECMWF	2	1 TB	2640 Power5+ (1.9 GHz)	20.1	2250 GB
FZJ	2	4 TB	1288 Power4 (1.7 GHz)	8.9	5152 GB
IDRIS	2	2 TB	1024 Power4 (1.3 GHz)	6.7	3136 GB
LRZ	(RZG)	0 TB	9728 Montecito (1.6 GHz)	62.3	39064 GB
RZG	2	10 TB	928 Power4 (1.3 GHz)	4.6	2368 GB

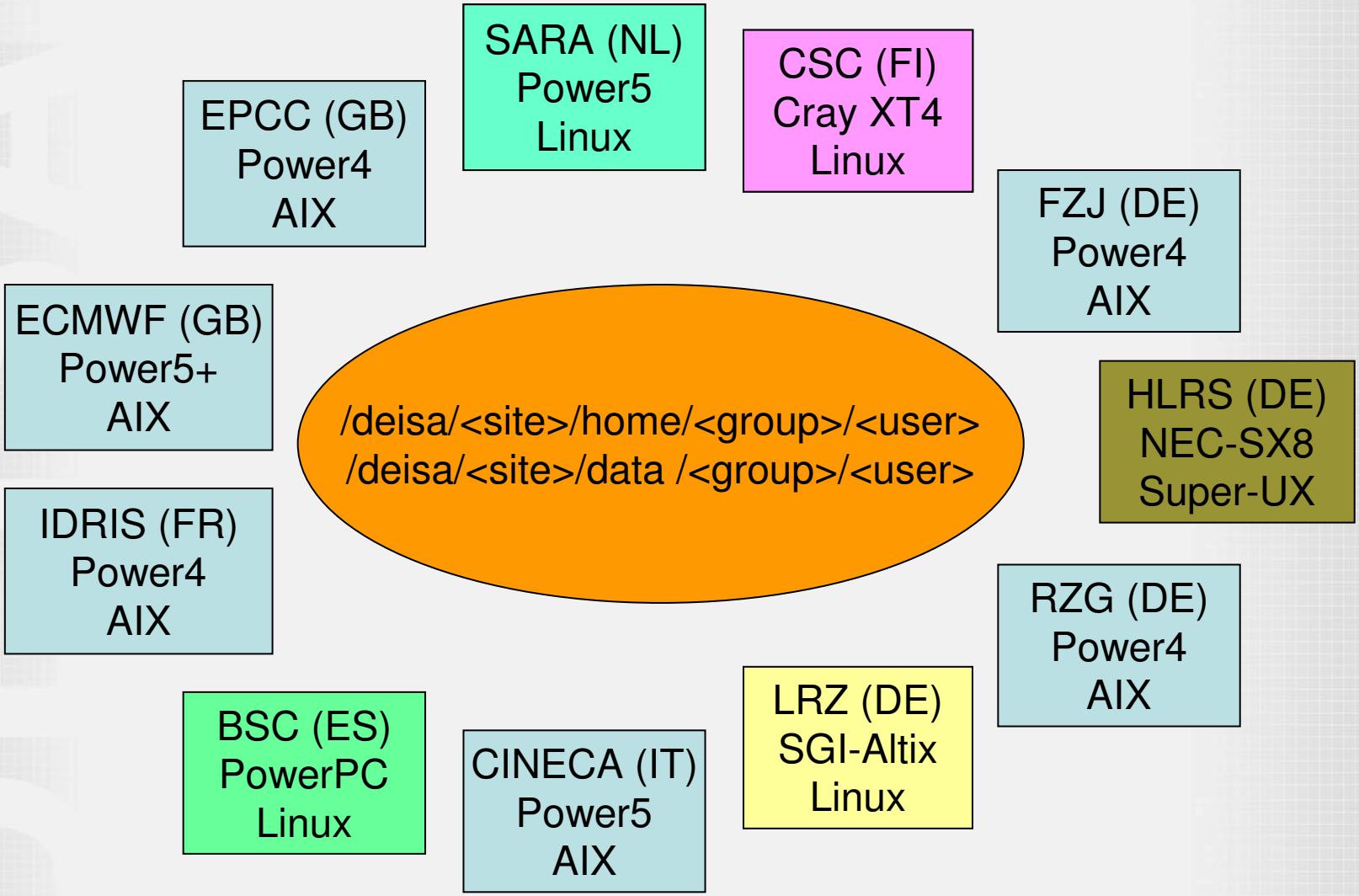
DEISA – Network (estimated Q3 / 2007)

Distributed European Infrastructure for Supercomputing Applications



ralph.niederberger@fz-juelich.de

Evolution of GPFS in DEISA



Discussion

Distributed European Infrastructure for Supercomputing Applications

Questions?